

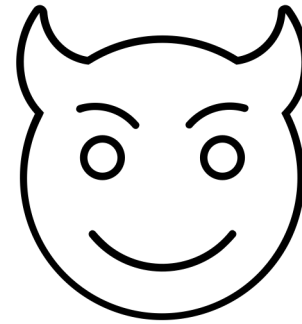
AI safety is essential for future society

2024-06-14

AI Safety Institute (AISI)

Is AI friend or not?

- ◆ If you use AI properly, you can achieve great results.
- ◆ However, if you use AI without understanding it, you may get involved in unexpected problems.



- ◆ We need knowledge and social mechanism.

Innovation Strategy 2024

- ◆ AI safety is an essential requirement for an environment that encourages innovation.

① AI innovation and the acceleration of innovation through AI

- Strengthening R&D capabilities (including data supply)
- Acceleration of the use of AI
- Enhancing AI infrastructure
- Human resource development and recruitment

② Realize the AI Safety

- Governance and rules
- AI safety
- Prevention of mis/dis information
- Intellectual property rights

③ International cooperation/collaboration

Overview of the AI Safety Institute (AISI)

Background

- ◆ October 2023
 - Agreed to the Hiroshima AI Process "International Guiding Principles" and "International Code of Conduct"
- ◆ November 2023
 - AI Safety Summit hosted by the U.K.
- ◆ December 2023
 - Agreement on "Hiroshima AI Process Comprehensive Policy Framework"
 - Prime Minister Kishida Announced Establishment of AISI
- ◆ **February 14, 2024**
 - AI Safety Institute (AISI) was established

About AISI

◆ Objectives

- **AISI supports public and private sector efforts.**
 - The public and private sectors need to work together to ensure that all parties involved in the development and use of AI are properly aware of the risks of AI. Governance also needs to be ensured throughout the lifecycle. Then, the safe and secure use of AI will be promoted.
 - Need to promote innovation and mitigate risks in the lifecycle, in those efforts.

◆ Principles

- **AISI activities to be harmonized with related domestic and international organizations.**
 - Response to rapidly and globally advancing technologies.

Role and Scope of AISI

◆ Role

- **AISI supports the government** by conducting surveys on AI safety, examining evaluation methods, and creating standards.
- **As a hub for AI safety in Japan**, AISI will consolidate the latest information in industry and academia, and promote collaboration among related companies and organizations.
- **Collaborate with AI safety-related organizations.**
 - AISI is not an R&D organization.

◆ Scope

- **Set the scope flexibly** in the following AI related issues, while considering **global trends**.

<ul style="list-style-type: none"> • Social Impact • governance 	<ul style="list-style-type: none"> • AI System • contents 	<ul style="list-style-type: none"> • data
---	---	--

Business

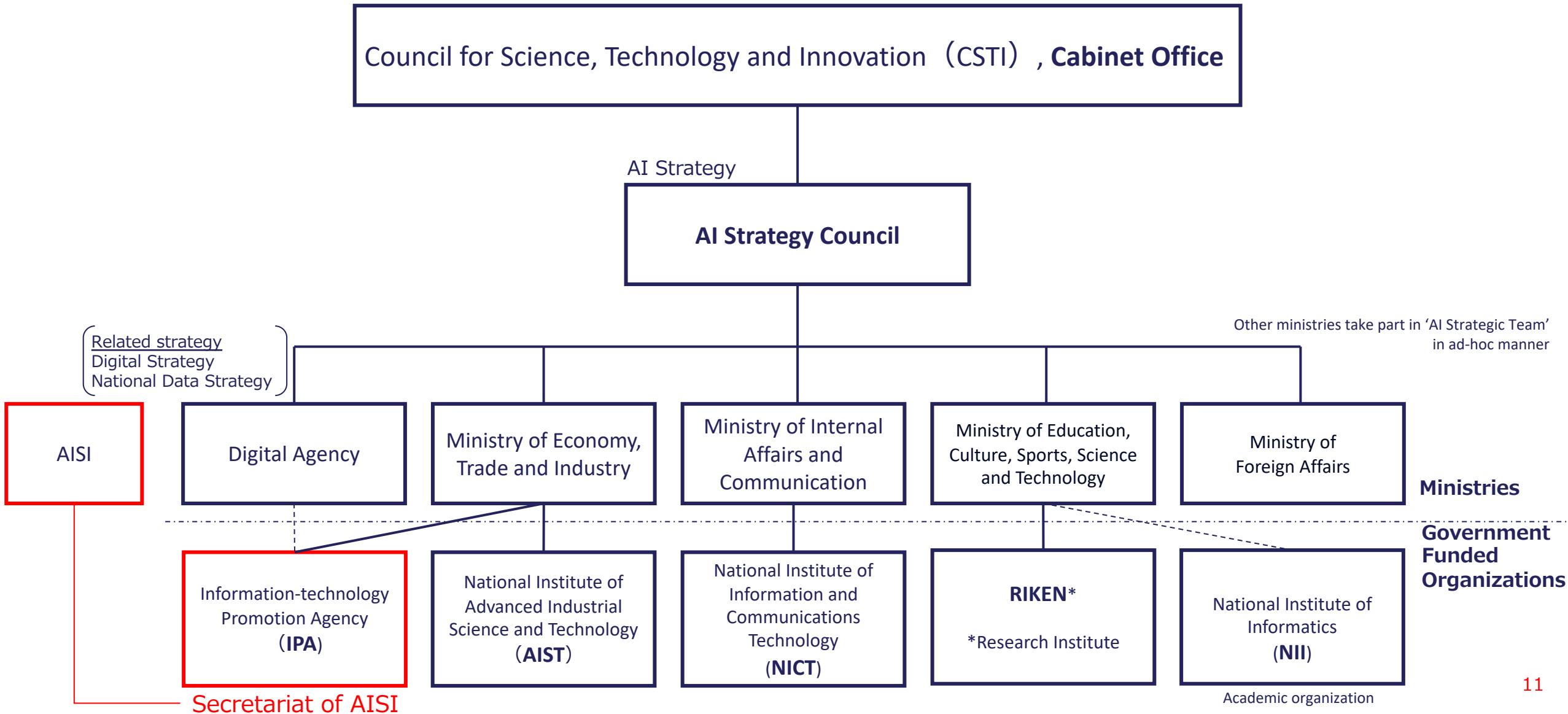
- 1) **Consideration of surveys and standards** for AI safety assessment
 - (i) Surveys on standards of AI safety, checking tools, anti-disinformation technology, AI and cybersecurity
 - (ii) Consideration of **standards and guidance** related to AI safety
 - (iii) Consideration of a testbed **environment** for AI related to the above
- 2) **Consideration of implementation methods** for AI safety assessment
- 3) **International collaboration** with related organizations in other countries (such as the AI Safety Institute in the U.K. and the U.S.)

Immediate Activities and Deliverables



Our team

AI Related Government organization



Executive Team

Executive Director
Akiko Murakami



Deputy Executive Director
Secretary General
Kenji Hiramoto

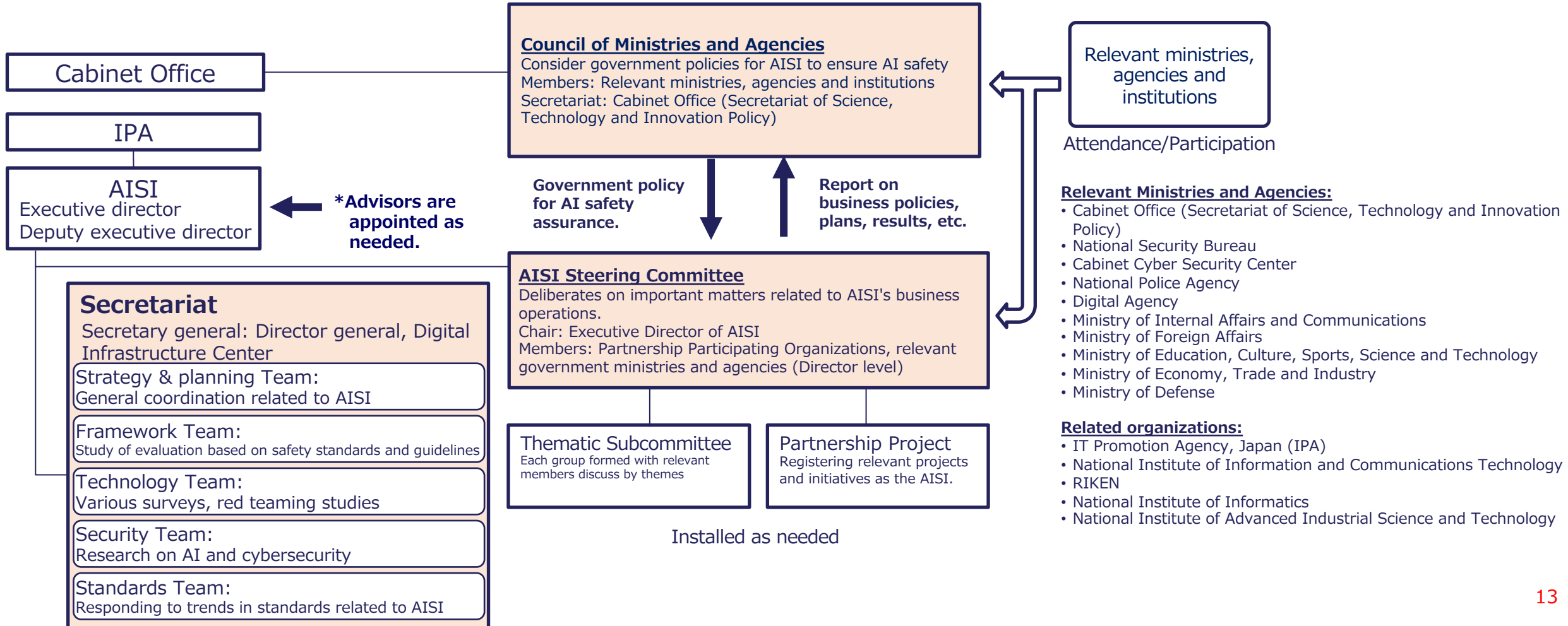


Deputy Executive Director
Hideyuki Teraoka



AISI Structures

- ◆ “Council of Ministries and Agencies”, set up in Cabinet Office, deliberates on the important matters of AISI.
- ◆ The “AISI Steering Committee” within AISI reports to the Council (to be held once a month). Under the Steering Committee, "thematic subcommittees" and "partnership projects" will be installed as necessary.
- ◆ As the secretariat of AISI, five teams were formed within the IPA Digital Infrastructure Center.



Items to be implemented at AISI

◆ Strategy & Planning

- Create the strategies and plans, Manage budgets
- Understand the situation regarding AI safety
- Public relations (e.g., AISI web site)
- Recruitment and human resources development
- Coordination and support with related organizations

◆ Framework

- Support for the development of AI guidelines (e.g., **Crosswalk**)
- Support for international coordination of AI risk management frameworks
- Collection of information related to AI governance and Technical Advice
- Support for consideration of the nature of **certification and accreditation**

Items to be implemented at AISI

◆ Technology

- Organizing **Red Team** Implementation Methodology
- Collect information and provide advice that contributes to the development of technology-related standards, guidelines, etc.
 - **Synthetic content, disinformation and misinformation**
 - Bias, data checking, history management
- Consideration of tools such as test beds, etc.

◆ Security

- Consideration of security measures for AI
- AI-based support for countermeasures against security events

◆ Standards

- Support for the promotion of ISO SC42
- Collection of other standard information

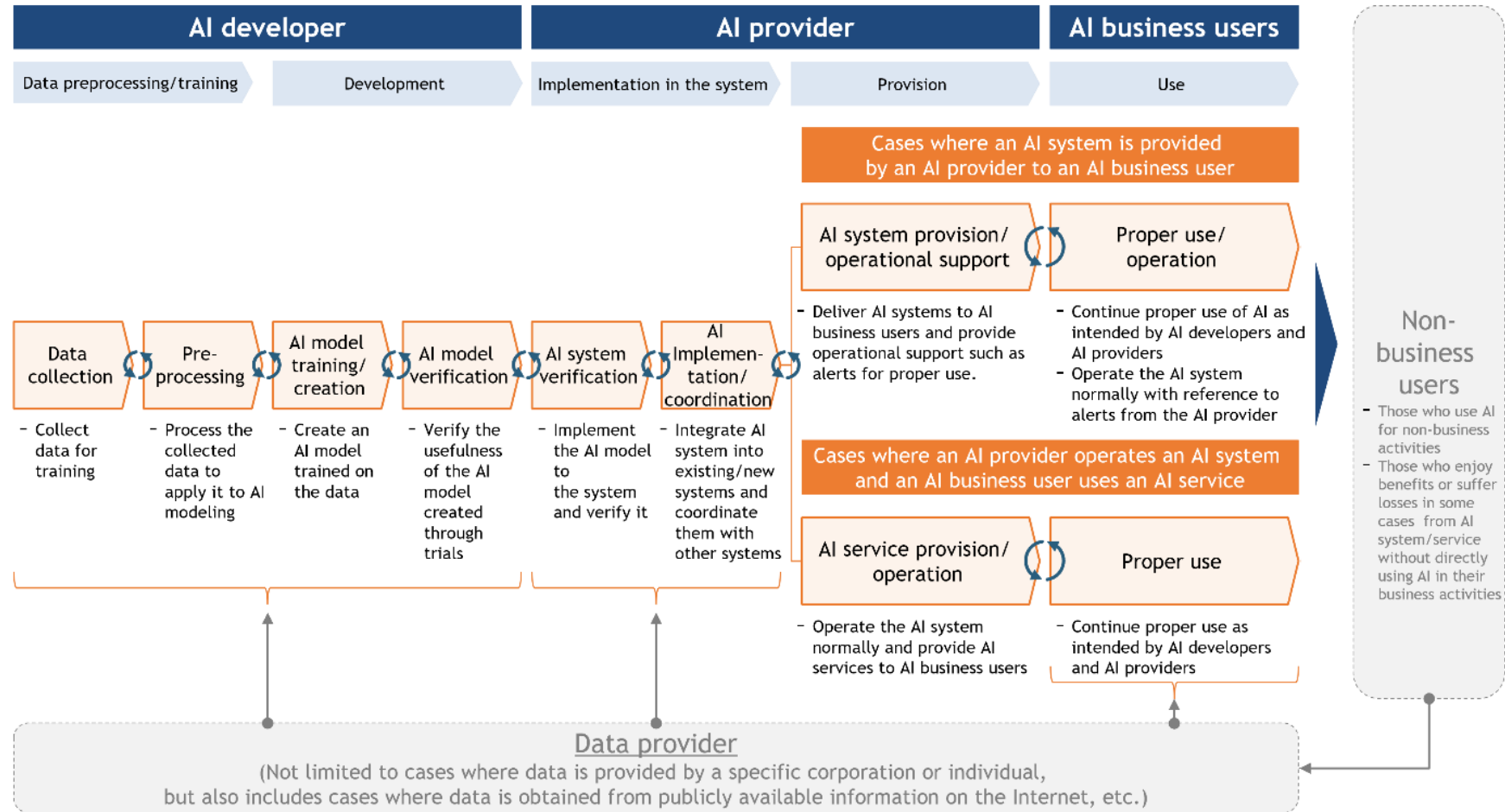
AI safety is essential for future society



Appendix: Recent developments related to AISI activities

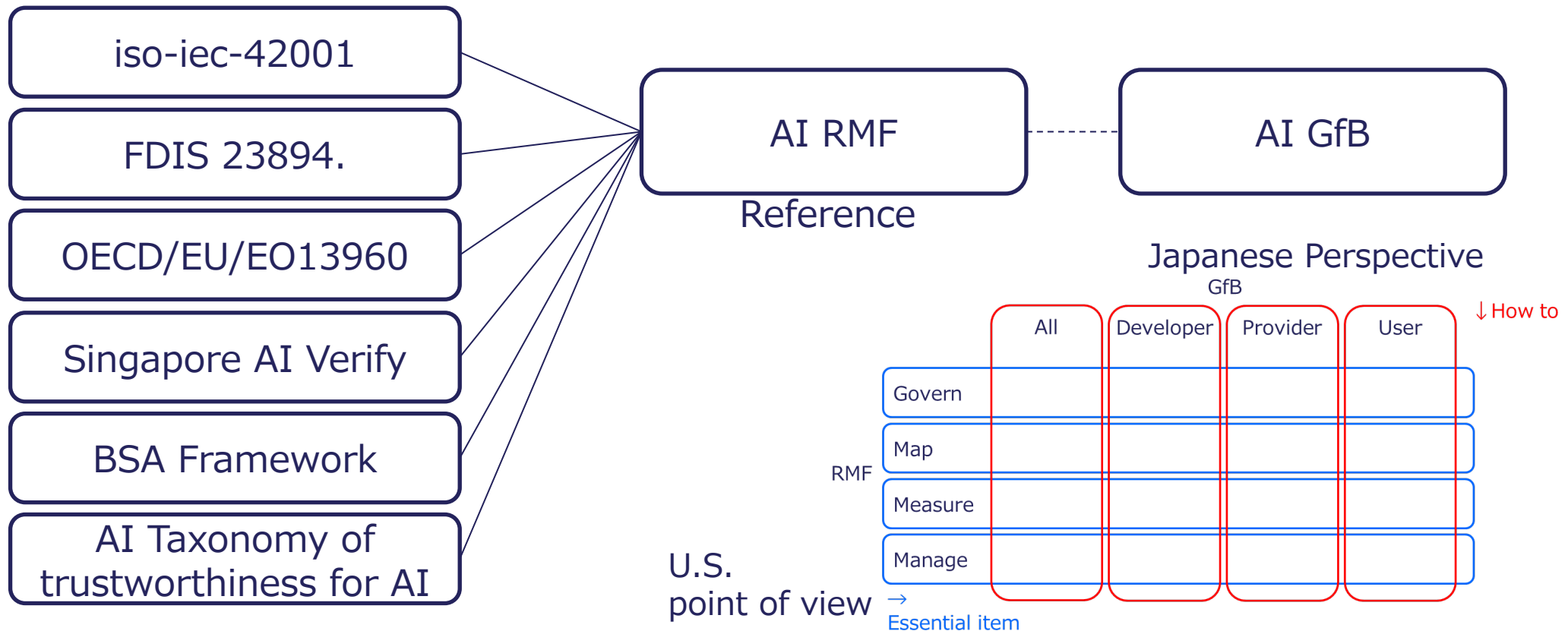
AI Guidelines for Business

- ◆ Clarify what each stakeholder should address in the flow of AI utilization



Japan-U.S. Crosswalk

- ◆ Confirmation of the interrelationship between the U.S. NIST AI Risk Management Framework (RMF) and the Japanese AI Guidelines for Business (GfB)



AISI

Japan AI Safety Institute